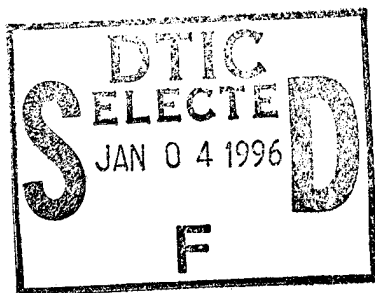


NAVAL HEALTH RESEARCH CENTER

REPLICATION AND EXTENSION OF MEASUREMENT MODELS FOR MANIKIN TEST, STERNBERG MEMORY TEST, AND PATTERN RECOGNITION TEST



*R. R. Vickers, Jr.
R. Hilbert
J. A. Hodgdon
R. L. Hessink
A. C. Hackney*

DTIC QUALITY INSPECTED 2

Report No. 92-13

Approved for public release: distribution unlimited.



NAVAL HEALTH RESEARCH CENTER
P. O. BOX 85122
SAN DIEGO, CALIFORNIA 92186 - 5122

NAVAL MEDICAL RESEARCH AND DEVELOPMENT COMMAND
BETHESDA, MARYLAND

19960102 008

Replication and Extension of Measurement Models for
Manikin Test, Sternberg Memory Test, and Pattern Recognition Test*

Ross R. Vickers, Jr.¹
Raymond Hilbert²
James A. Hodgdon³
Robert L. Hesslink³
Anthony C. Hackney⁴

Cognitive Performance and Psychophysiology Department¹,
Information Systems Division²,
and
Physiological Performance and Operational Medicine Department³

Naval Health Research Center
P.O. Box 85122
San Diego, CA 92186-5122

and

Department of Exercise and Sports Science⁴
University of North Carolina
Chapel Hill, NC 27599

Accession For	
NTIS	CRA21 <input checked="" type="checkbox"/>
DTIC	TAB <input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution /	
Availability	
Dist	Avail
A-1	

*Technical Report 92-13 supported by the Naval Medical Research and Development Command, Navy Medical Command, Department of the Navy, under Research Work Unit MM33B30.001-6104. The views presented are those of the authors and do not reflect the official policy of the Department of the Navy, Department of Defense, nor the U.S. Government.

Summary

Portable computerized cognitive assessment batteries provide a technology for standardized assessment of cognitive functioning in many applied settings where such assessment would have been impossible without this technology. The available scientific evidence guiding the interpretation of findings using these tests derives primarily from laboratory studies. Some of the procedures used in the laboratory are difficult to employ in the field because of the need to minimize interference with ongoing work. It is important to determine whether the relaxation of procedural controls needed to make field studies feasible significantly alters the interpretation of the measurements obtained. An earlier study indicated that the recommended laboratory procedure of including a series of familiarization trials could be relaxed without significant loss in the psychometric characteristics of the resulting measures. The present study attempted to replicate and extend this prior finding.

Study participants ($n = 36$) were male Norwegian military personnel participating in cold weather training. These men performed a computerized cognitive assessment battery consisting of a pattern recognition task, memory task, and spatial orientation task six times to obtain baseline measurement before undergoing studies of the effects of military training exercises on performance readiness. One extension of the prior work was that several alternative performance measures were considered for each cognitive test, rather than a single assessment. The measures typically included the number of problems attempted, the number correct, number of errors, percentage correct, average reaction time for correct responses, and the standard deviation of the reaction time for correct responses. Prior analyses had been limited to a single performance measure (e.g., reaction time) or a single composite (number right minus number wrong).

Performance scores were analyzed by structural equation modeling procedures. The basic structural model assumed that performance had constant true score variance and constant error variance across test sessions. Alternative models were constructed to correct for misfits between the data and this initial model. A second extension of prior work involved the cross-validation of these structural models by determining whether models could be constructed which fit the data from this sample and from a sample analyzed in a previous study.

A prior finding that performance generally involved constant true score variance from the first session onward was replicated. A tendency for error variance to be higher for the first

testing session than for later testing sessions also was replicated. The two samples produced generally consistent estimates of error variance after the first testing session for all three measures, but true score variance was comparable only for spatial orientation and memory.

Combined with the results of the previous study, the present findings indicate that a single familiarization trial is adequate for field studies. The replication of measurement models across two independent samples is consistent with the possibility of constructing normative models for cognitive performance in military personnel tested under neutral conditions to replace the use of control groups in field studies. The error variance estimates can be used to conduct statistical power analyses to ensure optimal sample sizes. Applying these findings can increase research efficiency by decreasing the number of subjects required for studies and abbreviating the testing time required for each subject in field studies of cognitive performance.

Introduction

Background

Portable computerized cognitive testing batteries now make it possible to evaluate cognitive performance capabilities by standardized procedures administered under demanding conditions in real-life settings that could not be readily studied even a few years ago. A substantial body of prior research which describes the development of cognitive tests for use in laboratory settings provides an empirical track record for such tests with respect to psychometrics and sensitivity to experimental conditions (AGARD Working Group 12, 1989; Bittner, Carter, Kennedy, Harbeson, & Krause, 1986; Perez, Masline, Ramsey, & Urban, 1987). This prior track record derives largely from work performed in research laboratories using procedures appropriate to that setting. Applications of these tests to field research where there typically is less control over the environment and less time available for testing presents special problems that may require modifications of previous procedures. Such modifications raise issues concerning the generalizability of the results obtained in the laboratory to field settings. These issues must be investigated to verify that laboratory-based tests can be appropriately interpreted when data are gathered in the field.

One element of typical laboratory research designs that may require modification under field conditions is the use of training sessions prior to collection of data to assess effects of exposure to some experimental condition. Research protocols often may have to be adapted to time slots defined by the ongoing job requirements of the people studied. Job requirements often will limit the time available for testing, thereby making it necessary to truncate or completely eliminate practice sessions used to establish baselines in the laboratory.

The potential need to drop or truncate baseline performance assessments faces the researcher with a dilemma. The baseline measures are intended to ensure that results are interpretable and meet basic statistical requirements for significance testing (Bittner et al., 1986; Carter & Woldstad, 1985; Jones, Kennedy, & Bittner, 1981). The research design dilemma involves a choice between employing procedures which provide less than ideal bases for assessing the effects of interest and not doing the study at all. Obviously, if data quality would be unacceptably low for the designs that are feasible within the research constraints, research

interests would be served better by not conducting the study. Faced with this dilemma, the research design problem is to decide whether the data which can be obtained within the temporal constraints imposed for a particular study will be of sufficient quality to permit legitimate inferences to be drawn from the study.

The dilemma facing the field researcher may be less severe than it appears. Vickers and Kusulas (1992) reanalyzed data from Kennedy, Dunlap, Jones, Lane, and Wilkes (1985) and found that the typical test in a set of cognitive measures has constant true score variance from the first session onward. The major problem identified in their analyses was low measurement precision of the initial session with a test. If these findings can be replicated, the research dilemma defined above can be circumvented by research protocols that employ designs suitable for controlling learning curve effects and samples large enough to overcome the measurement limitations of the first trial (AGARD Working Group 12, 1989; Vickers & Kusulas, 1992). Protocols could be substantially simplified if experimental design concerns could be limited to controlling for learning curve effects, so the present study attempted to replicate the prior findings in a different population.

Research Issues

The conclusions reached by Vickers and Kusulas (1992) must be viewed with caution. These conclusions were derived from a data base which involved a single, relatively small sample and considered a limited range of cognitive tests relative to the alternatives available in batteries now in use (AGARD Working Group 12, 1989; Englund et al., 1985).

The present study extended the prior work in three ways. First, an attempt was made to replicate the previous finding that cognitive battery tests provide at least tau-equivalent measurement from the first testing session. Tau-equivalent measurement means that a test has constant true score variance over a series of sessions, but the magnitude of random error variance changes (Lord & Novick, 1968). In this particular instance, the previous evidence indicated that error variance was large for the first testing session, then relatively constant.

The second extension of the prior work was a more stringent test of the replicability of measurement models than simply determining whether a tau-equivalent model was appropriate for a given cognitive test. In this extension, parameter estimates for true score variance and error variance derived in the prior study were applied to data from the present study. If these estimates

fit the data well relative to values derived directly from these present data, the models receive stronger confirmation than if different parameter values must be estimated for each sample. The confirmation of the model is stronger under these conditions because the range of alternative parameter values which are regarded as providing confirmation of the model is more restricted than it was in the initial study. In the initial study, any set of non-zero parameter values consistent with the general form of the structural equation model which reproduced the covariance matrices was considered acceptable confirmation of these models. In the present replication, the range of acceptable values is constricted to those values equal to the earlier parameter values plus or minus sampling error. The specification of point estimates for the structural parameters rather than the general hypothesis that these parameter values are not equal to zero implies a narrower range of alternatives and, in that sense, a stronger test of the models (Meehl, 1991).

The present study applied statistical procedures for estimating the replicability of model parameters to data for the Sternberg Memory test (Sternberg, 1966) and the Manikin test (Benson & Gedye, 1983). This extension of the prior work was undertaken because one programmatic objective for this project is to develop statistical models for use as baseline reference points for testing the effects of environmental stressors. If replicable models can be developed under neutral conditions, study designs which employ only an experimental group are feasible. If not, experimental and control group designs probably will be necessary to derive strong inferences from field studies.

The third extension of the prior work examined the effect of choosing different performance measures from those which were available for a given test. The previous study relied on published data, so analyses were limited to just those aspects of performance reported for each test (e.g., number correct, net correct minus incorrect, or reaction time). The present study worked with raw data which included multiple possible measures of performance, such as number correct, number incorrect, and response times. These different aspects of performance may assess different, relatively independent components of the cognitive structures that determine overall performance (e.g., Baddeley, 1986). It was desirable, therefore, to have suitable measurement models for all of the components of performance.

The issues outlined above were addressed by analyzing cognitive test performance of military personnel undergoing cold weather training. The analyses employed cognitive test scores from six familiarization sessions conducted prior to beginning the actual training. The familiarization sessions were conducted as part of the typical procedure for establishing baseline performance prior to testing for effects of exposure to factors which might modify performance. The structural modeling procedures employed by Vickers & Kusulas (1992) were applied to this data with modifications incorporated to address the topics indicated above.

Method

Sample

Study participants were 36 Norwegian Army personnel who volunteered to participate in a study of the effects of cold weather on cognitive and physical performance. All volunteers were male Caucasians, and all had completed the Norwegian equivalent of high school. The average age of the participants was 20.6 years ($SD = 1.4$; range = 19 - 27).

Cognitive Tests

The cognitive tests used in this study were taken from the Essex Cognitive Test Battery (Kennedy et al., 1985). The specific tests chosen were:

Pattern Recognition (Klein & Armitage, 1979). Subjects are presented two side-by-side patterns of asterisks and asked to indicate whether they are the same or different. The patterns involve eight dots located at different points on the screen in a 4 x 4 grid. The subject is required to indicate whether the patterns are the same or different. The performance measures derived from this test included the number of pattern comparisons attempted, the number of correct comparisons, the reaction time for correct comparisons, and the standard deviation for the correct comparisons.

Sternberg Memory Test (Sternberg, 1966). A set of four target letters is presented to the subject on the video screen. The subject is permitted as much time as he wants to memorize them. Probe letters then are presented, and the respondent is asked to indicate whether the probe is one of the target letters. Performance measures for this test that were utilized in the analyses included number of trials, number correct, average reaction time for correct responses, and standard deviation of response time for correct responses.

Manikin Test (Benson & Gedye, 1983). The Manikin test asks participants to determine whether a manikin figure is holding an object in his right hand or his left hand. On each trial the figure is presented holding two objects, one in each hand. The figure is standing on a box that contains one of the objects and the figure may be facing toward the

respondent or away from him. Direction of orientation can be determined by the presence or absence of facial features. The respondent must examine the box that the figure is standing on to determine which object is the target object, then determine which direction the manikin is facing and whether the hand holding the object of concern is on the subject's right or the subject's left. These three pieces of information then are combined to determine which of the manikin's hands is holding the object shown in the box. Measures used in the analyses included the number of attempts, the number of correct responses, the reaction time for correct responses, and the reaction time for incorrect responses.

Analysis Procedures.

The analysis strategy for testing the structural models began with the assumption that each test provided a series of parallel measures of individual differences across the six sessions. For this model to be correct, it would be necessary to demonstrate that true score variance was constant across sessions, that error variance was constant across sessions, and that scores on different sessions were not correlated after taking into account the constant true score variance. The degree to which these conditions were met by the present performance assessments was determined by constructing a structural equation model which embodied the following assumptions:

- (a) The scores for each session reflected differences on a single latent construct.
- (b) The magnitude of true individual differences in performance was constant across sessions, so the variance of these true scores was constant.
- (c) The frequency and magnitude of effects due to random factors was constant across trials.

Basic Model Specifications. The parallel tests assumptions were operationalized by imposing three constraints on the parameters used to reproduce the covariance matrices for the performance measures. The model included only a single latent trait (assumption (a)) and constrained scores from different sessions to have equal factor loadings (i.e., equal lambdas; assumption (b)) and equal error variance (i.e., equal epsilons; assumption (c)) across sessions. The assumption that a single trait was measured implies conditional independence of the scores from different sessions after the covariation arising from the single underlying trait has been taken into account (Glymour, Scheines, Spirtes, & Kelly, 1987; Kivii & Speed, 1982). This assumption was modeled by adding the constraint that residual covariances between scores from

different sessions were zero after taking into account covariation produced by individual differences in the hypothesized latent trait. In the following presentation and discussion of results, the last three constraints embodied in the structural equation model are referred to as the "equal true variance," "equal error variance," and "uncorrelated errors" constraints, respectively.

Model Modification Procedures. After the initial model was fitted to the data, modification indices provided by LISREL VI (Joreskog & Sorbom, 1981) were examined to identify important areas of misfit between the model and the present data. Modification indices indicate the improvement in fit of the model expected by removing the constraints on a specific parameter. Examples of how constraints might be relaxed would include letting the factor loading (λ) or the error variance (ϵ) for the first trial be different than the corresponding parameters for later trials. The constraint which would most improve the fit of the model then was removed, new estimates of λ and ϵ were computed. The fit of the resulting model was examined to see whether relaxing additional constraints would further improve the fit between the model and the data being analyzed. If so, the process was repeated until a final model was achieved.

Goodness-of-Fit Criterion. The criterion for stopping model modifications was based on incremental goodness-of-fit for alternative models. In general, goodness-of-fit indices indicate how well a given structural equation model reproduces the covariance matrix that is being analyzed. In the present study, Tucker and Lewis's (1973) index (hereafter, TLI) was chosen to provide the basic quantitative measure of fit between the model and the data. This index is based on the difference between the raw covariances being analyzed and the residual covariances after fitting a particular structural model. This difference is compared to the maximum difference that would be possible, i.e., the difference between the observed covariances and the expected magnitude of chance covariation. Thus, the measure corresponds to common methods of assessing proportional reduction in error as a means of comparing alternative models.

Typical evaluations of proportional reductions in error make allowances for the number of degrees of freedom employed to achieve that reduction. Parsimony adjustments have been recommended for evaluating structural equation models (Mulaik et al., 1989), so these adjustments were employed. The use of parsimony as a basis for choosing between models can be justified on philosophical grounds (Mulaik et al., 1989) and on the basis of improved precision

in estimating parameter values for models (Bentler & Mooijart, 1989). Applying these considerations, the model adopted to represent each performance measures was that which had the largest parsimony-adjusted TLI value (hereafter, ATLI).

Cross-Validation Procedures. The direct cross-validation of structural models was accomplished by applying the parameter estimates from Vickers and Kusulas (1992) to the present data. In practice, this aspect of the analysis involved reviewing the previous analyses to determine the parameter values, then fixing those parameters at those specific values instead of estimating them from the performance measures collected in this study. The goodness-of-fit of the resulting model indicates how well the model developed from that prior data reproduced the pattern of associations in the present data. Conceptually, this fit assessment is equivalent to the multiple correlation obtained when a regression equation is cross-validated.

Results

Parallel Tests Model

The parallel tests model provided reasonably good overall fit to the data considered across the 15 performance scores evaluated (Table 1). Eleven of the 15 models evaluated produced parsimony-adjusted fit indices of .70 or greater; three others were between .60 and .70. The only instance of extremely poor fit was the percent correct for the Sternberg Memory Test (ATLI = .37).

Sources of Misfit

When aggregated across all of the performance measures, the parameter constraints which produced the misfit between the model and the data tended to localize in the equal true score constraints and the equal error constraints for early session (Table 2). The exception to this general trend was the high average level of misfit for the equal true score constraint on session 6. The results for the uncorrelated errors constraint deviated from this general picture because this constraint had roughly equal effects across sessions except for session 4.

Table 1
Summary of Parallel Model Tests

	Chi-Square for:			
	<u>Null</u>	<u>Parallel</u>	<u>TLI</u>	<u>ATLI</u>
Manikin	427.29	134.36	.78	.76
# Correct	162.74	62.41	.69	.65
# Errors	232.93	75.53	.72	.69
% Correct	227.09	70.08	.74	.70
Avg. Response Latency	159.94	37.54	.86	.82
S.D. Response Latency	162.77	69.39	.63	.60
Sternberg Memory	296.29	83.75	.87	.85
# Correct	205.90	25.90	.96	.91
# Errors	86.33	30.23	.82	.78
% Correct	137.43	87.08	.39	.37
Avg. Response Latency	233.68	40.73	.89	.85
S.D. Response Latency	150.13	40.08	.83	.79
Pattern Recognition	432.55	103.48	.86	.85
# Correct	235.58	41.30	.89	.85
# Errors	96.01	28.15	.87	.83
% Correct	78.28	23.24	.92	.88
Avg. Response Latency	255.25	37.93	.92	.87
S.D. Response Latency	165.27	52.10	.76	.72

NOTE: The Null Model had 20 degrees of freedom and the Parallel model had 19 degrees of freedom. "TLI" refers to the raw Tucker-Lewis index. "ATLI" refers to the parsimony-adjusted Tucker-Lewis Index.

Table 2
Summary of Cumulative Modification Indices for Different Sessions

Trial	Cumulative Chi-Square for Constraint of:					
	Equal True Score		Equal Error		Uncorrelated Error	
	Average	Maximum	Average	Maximum	Average	Maximum
1	2.44**	11.46	8.65**	61.29		
2	3.86**	32.42	4.19*	30.78	3.44**	11.08
3	1.98*	6.49	1.89*	4.77	2.94**	11.71
4	1.59	4.70	2.65**	8.83	.52	1.71
5	1.67*	5.93	2.24*	5.96	3.63**	13.06
6	4.25**	11.91	2.43**	17.86	4.41**	16.40

* $p < .05$ (Critical value = 1.67, 15 df)

** $p < .0033$ (Bonferroni critical value = 2.27, 15 df)

Model Modifications

The stepwise modification procedures used to derive alternatives to the parallel tests model involved 21 modifications for the 15 performance measures (Table 3). Several findings were consistent with trends in the results reported by Vickers and Kusulas (1992), including:

- (a) The level of fit to the data provided by the parallel tests model was comparable despite a somewhat poorer fit in the present study. The median ATLI in the present study was .785 (range = .37 - .91) compared to .845 (range = .28 - .91) for eight models in the prior study.
- (b) Modifications of the equal error constraint (10) were more common than modifications of the uncorrelated error constraint (6) and the equal lambda constraint (5).
- (c) The modifications of different constraints were not randomly distributed across the cognitive tests. Equal lambda constraint modifications were found primarily for the Manikin test (4 of 5), correlated error constraint modifications were found primarily for the Pattern Recognition test (5 of 6), and equal error constraint modifications were located predominantly in the Sternberg Memory Test models (6 of 10).
- (d) Overall, 5 of 15 performance measures produced parallel tests as the final model.
- (e) Violations of the equal error constraint were not randomly distributed across sessions as 5 of 10 such modifications were for session 1.

Trends that could not be anticipated from the prior study included the following:

- (a) Model modifications were evenly distributed across the early sessions (11) and late sessions (10) despite the general tendency toward larger chi-squares for early sessions (cf., Table 2). In the prior study, 5 of 7 modifications were for sessions 1 or 2.
- (b) Minor changes in the stopping rule for model modification would have led to the adoption of the parallel test for 10 of 15 performance measures. If the criterion had been an improvement of more than .01 in the ATLI, the parallel model would have been adopted for 5 additional models: (i) Pattern Recognition # Correct, (ii) Pattern Recognition # Errors, (iii) Pattern Recognition % Correct; (iv) Manikin SD Response Latency, and (v) Sternberg Memory # Correct.
- (c) The parallel tests model tended to fit best for average response latency measures. While 5 of 15 performance measures were fitted by this model, all three models for average response latency were fitted by this model.

Table 3
Summary of Model Modifications

	Chi-Square for: Improvement				Chi-Square
	<u>Model</u>	<u>in Fit</u>	<u>TLI</u>	<u>ATLI</u>	<u>%</u>
Pattern Recognition					
# Correct					
(a) Parallel	62.41	100.33	.68	.65	.71
(b) Error 1	52.43	9.98	.73	.66	.78
(c) Err Corr 21	43.97	8.46	.78	.66	.84
# Errors					
(a) Parallel	75.53	157.40	.72	.68	.76
(b) Err Corr 56	62.91	12.62	.77	.69	.83
(c) Err Corr 32	50.89	6.89	.81	.69	.88
% Correct					
(a) Parallel	70.08	157.01	.74	.70	.76
(b) Err Corr 56	57.78	12.30	.79	.71	.82
(c) Err Corr 32	45.85	11.93	.84	.71	.88
Avg. Response Latency					
(a) Parallel	37.54	122.40	.86	.82	.84
S.D. Response Latency					
(a) Parallel	69.39	93.38	.63	.60	.64
(b) Lambda 2	32.72	36.67	.89	.80	.90
(c) Error 1	24.17	8.55	.94	.80	.95
Manikin					
# Correct					
(a) Parallel	25.90	180.00	.96	.91	.96
# Errors					
(a) Parallel	30.23	56.10	.82	.78	.82
(b) Error 1	21.32	8.91	.94	.85	.94
% Correct					
(a) Parallel	87.08	50.35	.39	.37	.41
(b) Error 1	47.84	39.24	.72	.65	.74
(c) Lambda 6	37.17	10.57	.80	.68	.82
(d) Lambda 5	29.92	7.25	.85	.68	.88
(e) Lambda 4	15.72	14.20	.99	.74	1.00
Avg. Response Latency					
(a) Parallel	40.73	192.95	.89	.85	.90
S.D. Response Latency					
(a) Parallel	40.08	110.05	.83	.79	.81
(b) Lambda 6	30.62	9.46	.89	.80	.88

Table 3
Summary of Model Modifications
(Continued)

	Chi-Square for: Improvement in Fit		Chi-Square		
<u>Model</u>			<u>TLI</u>	<u>ATLI</u>	<u>%</u>
Sternberg Memory					
# Correct					
(a) Parallel	41.30	194.28	.89	.85	.91
(b) Error 6	28.06	13.24	.95	.85	.97
# Errors					
(a) Parallel	28.15	67.86	.87	.83	.80
(b) Error 1	22.89	5.26	.93	.84	.87
(c) Error 4	16.83	6.06	1.00	.85	.94
(d) Error 6	11.55	5.28	1.07	.86	1.00
% Correct					
(a) Parallel	23.24	55.04	.92	.88	.87
Avg. Response Latency					
(a) Parallel	37.93	217.32	.92	.87	.91
S.D. Response Latency					
(a) Parallel	52.10	113.17	.76	.72	.77
(b) Error 2	43.09	9.01	.81	.73	.84
(c) Corr Err 32	33.41	9.68	.87	.74	.90
(d) Error 4	19.11	14.30	.97	.78	1.00

NOTE: "Parallel" = parallel tests model; "Lambda," "Error," and "Err Corr" = type of constraint freed; Numbers = testing session(s). Thus, "Error 1" refers to a model produced from the prior model by removing the equal error constraint for the first session. "Improvement in fit" is the chi-square change from the prior model; the null model was used as the prior model for the Parallel model. "TLI" is the Tucker-Lewis index; "ATLI" is the parsimony-adjusted TLI. "Chi-Square %" = (Null Model Chi-Square - Current Model Chi-Square) / (Null Model Chi-Square - Minimum Chi-Square). "Minimum Chi-Square" is the chi-square obtained by continuing model modification until all modification indices were less than 3.84.

Revised General Model

The aggregate results of the two studies performed to date suggest that permitting the error magnitude for the first test session to differ from that for later sessions would provide an alternative model that might fit most of the available data. Introducing this modification typically improved the fit between the model and the covariance matrices relative to the fit obtained with the parallel tests model (Table 4). In the case of Pattern Recognition, the improvement in fit of the model was statistically significant for 2 of 5 performance measures. The cumulative

chi-square would be statistically significant if the results for the five Pattern recognition performance measures were regarded as independent (chi-square = 19.65, 5 df, $p < .002$). Application of the Revised General Model to the Manikin Test results produced significant improvement in fit for 3 of 5 performance measures and a cumulative chi-square of 55.22 (5 df, $p < .001$) for the five Manikin Test performance measures taken as a set. The comparable figures for the Sternberg Memory Test were significant improvement in fit for 3 of 5 performance measures with a cumulative chi-square of 19.53 (5 df, $p < .002$).

Table 4
Comparison of Parallel Tests Model and Revised General Model

<u>Test</u>	Chi-square Results:			Goodness-of-fit for:			
	<u>Parallel</u>	<u>Revised</u>	<u>Delta</u>	<u>Parallel</u> <u>TLI</u>	<u>ATLI</u>	<u>Revised</u> <u>TLI</u>	<u>ATLI</u>
<u>Pattern Recognition</u>							
# Correct	62.41	52.43	9.96	.68	.65	.73	.66
# Errors	75.53	75.17	0.36	.72	.68	.70	.63
% Correct	70.08	70.05	0.03	.74	.70	.72	.65
Avg. Latency	37.54	29.55	7.99	.86	.82	.91	.82
S.D. Latency	69.39	68.10	1.29	.63	.60	.61	.55
<u>Manikin</u>							
# Correct	25.90	24.67	1.23	.96	.91	.96	.86
# Errors	30.23	21.32	8.91	.82	.78	.94	.85
% Correct	87.08	47.84	39.24	.39	.37	.72	.65
Avg. Latency	40.73	39.43	1.30	.89	.85	.89	.80
S.D. Latency	40.08	35.54	4.54	.83	.79	.85	.77
<u>Sternberg Memory</u>							
# Correct	41.30	40.09	1.21	.89	.85	.89	.80
# Errors	28.15	22.89	5.26	.87	.83	.93	.84
% Correct	23.24	15.48	7.76	.92	.88	1.05	.94
Avg. Latency	37.93	33.02	4.91	.92	.87	.93	.84
S.D. Latency	52.10	51.61	0.49	.76	.72	.74	.67

Combining the results for all three cognitive performance tests, the Revised General Model produced a statistically significant improvement in fit in 8 of 15 cases. Applying the binomial probability model, the probability that 8 results would exceed the $p < .05$ significance criterion in a series of 15 independent significance tests is $p < 10^{-6}$. The total chi-square improvement from the Parallel Tests Model to the Revised General Model was 94.8 (15 df, $p < .001$).

Examination of the change in goodness-of-fit indices indicated that the revised general model had less of an advantage over the parallel tests model than might have been inferred from the changes in the chi-square statistics. While 5 of 15 goodness-of-fit measures improved, 9 of 15 decreased. The average ATLI for the parallel and revised general models was closely comparable, but this similarity was substantially influenced by the much better fit of the general revised model for Manikin Test Percent Correct.

Replication for Specific Performance Measures

The replicability of measurement models was tested by creating performance composites comparable to those used by Kennedy et al. (1985). These performance measures were composites because they were comprised of the number correct (or number attempted) minus number wrong divided by time on task. The structure of these measures in the present sample then was compared to the structure in the Kennedy et al. (1985) sample.

Analysis Procedure. Multiple group analysis procedures were applied to estimate factor loadings and error terms for different structural models using the combined data from both samples simultaneously. The structural form of the model was specified, and parameter values were estimated with the initial constraint that these values be the same for both samples. Subsequent models tested the hypothesis that the two samples could be represented adequately by a single measurement model by removing the equality constraint on one or more of the model parameters. If removal of the constraint did not produce a significant improvement in the fit of the model to the data, the two groups could be regarded as equivalent with respect to that model parameter. The specific models tested in this fashion are outlined below.

Parallel Tests Model. The initial model was a parallel tests model which assumed that the true score variance and error variance were constant from the initial session onward and were equal in the two samples. This model provided reasonably good fit to the data (ATLI = .760 - .865; Table 5).

Table 5
Goodness-of-Fit Results for Cross-Validation

	<u>df</u>	<u>Chi-Square</u>	<u>TLI</u>	<u>ATLI</u>
<u>Pattern Recognition</u>				
Null	41	405.43		
Parallel	40	113.84	.792	.773
Unequal Lambdas	39	106.87	.804	.765
Unequal Epsilons	39	85.18	.860	.797
General Revised	38	79.19	.884	.841
Separate Parallel	38	78.95	.879	.814
<u>Sternberg Memory</u>				
Null	41	356.58		
Parallel	40	74.93	.887	.865
Unequal Lambdas	39	72.99	.887	.844
Unequal Epsilons	39	65.16	.907	.841
General Revised	38	74.85	.881	.838
Separate Parallel	38	63.43	.913	.846
<u>Pattern Recognition</u>				
Null	41	405.43		
Parallel	40	140.64	.779	.760
Unequal Lambdas	39	127.89	.800	.761
Unequal Epsilons	39	140.54	.763	.707
General Revised	38	137.33	.778	.740
Separate Parallel	38	127.75	.792	.734

Different Lambdas Model. The second model removed the initial constraint that the factor loadings be equal across the two samples while retaining the assumption that the error variance was constant. This model would be appropriate if the measurement precision of a test were constant for the two samples, but sampling variation altered true score variance in the cognitive abilities determining test performance. Relative to the Parallel Tests Model, removing the equal lambdas constraint produced significant changes in fit for the Manikin test (chi-square = 6.97, 1 df, $p < .009$) and Pattern Recognition (chi-square = 12.75, 1 df, $p < .001$), but not for Sternberg Memory (chi-square = 1.94, 1 df, $p < .164$). The ATLI for the Manikin test data was less than that for the Parallel Tests Model (.765 versus .773), so the overall findings indicated improved fit for Pattern Recognition, no improvement in fit for Sternberg Memory, and equivocal results for the Manikin test.

Different Errors Model. The third model tested the hypothesis that the precision of measurement differed between the Norwegian sample and the Kennedy et al. (1985) sample. In

this instance, the constraint on equal epsilons for the two samples was removed. The resulting change in the chi-square values was statistically significant for the Manikin test (chi-square = 28.66, 1 df, $p < .001$), but not for the Sternberg Memory test (chi-square = 3.04, 1 df, $p < .082$) or the Pattern Recognition test (chi-square = 0.10, 1 df, $p < .752$).

Revised General Model. The fourth model fitted to the data assumed that the Revised General Model could be applied with equal parameter values in each sample. This model produced significant improvement in fit relative to the parallel tests model for the Manikin test (chi-square = 34.65, 1 df, $p < .001$). However, the Revised General Model produced only slight, statistically nonsignificant, improvements in fit relative to the parallel tests model for the Sternberg Memory test (chi-square = 0.08, 1 df, $p < .778$) and the Pattern Recognition test (chi-square = 3.31, $p < .069$).

The fifth model fitted to the data assumed that a parallel tests model fit the data within each sample, but that both lambda and epsilon differed for the two samples. This model was compared to the best fitting of the prior models for each test. For the Manikin test, the assumption of separate parallel models did not improve significantly on the Revised General Model (chi-square = 0.24, 1 df, $p < .625$). For the Sternberg Memory test, the assumption of separate parallel tests did not improve significantly on the unequal epsilon model (chi-square = 1.73, 1 df, $p < .189$). For the Pattern Recognition Test, the assumption of separate parallel tests did not improve significantly on the unequal lambdas model (chi-square = 0.14, 1 df, $p < .709$).

Best Fitting Models. The largest goodness-of-fit index is one guide to which model to adopt from among those compared. By this criterion, the best fitting model was the Revised General Model for the Manikin test, the Parallel Tests model for the Sternberg Memory test, and the Different Lambdas model for the Pattern Recognition test. The magnitude of these indices were .841, .865, and .761, respectively. For Pattern Recognition, the Parallel Tests model produced a goodness-of-fit so close to that obtained with the Different Lambdas model (.760 versus .761) that either model could reasonably be adopted on the basis of the present evidence.

Location of Misfit between Cross-Validation Models and Data. Modification indices for the best fitting models were examined to identify model modifications which would improve the fit of the model to the data. The Revised General Model fit the Manikin Test data best. The modification indices for this model indicated substantial improvements in fit could be expected

if the constraint that the session 1 error component be equal for both samples was removed. Relaxing this constraint produced a model with a better overall fit to the data (chi-square = 62.84, chi-square change = 16.35, 1 df, $p < .001$), but the improvement in fit was not sufficient to produce a higher ATLI for the revised model (ATLI = .832 versus ATLI = .841 for the model with equal Error 1 values in both samples).

The prior analyses indicated that the parallel tests model provided the best fit for the Sternberg Memory test. The modification indices indicated that constraining the error for session 6 in the Norway data to be equal to the error on the other trials in this sample was the single greatest source of misfit. Permitting the value of this parameter to be estimated freely for the Norway sample substantially reduced the misfit between model and data (chi-square = 55.57; change in chi-square = 19.36, 1 df, $p < .001$). This reduction in misfit was enough to improve the ATLI (.926 versus .865).

The unequal lambdas model fit the data best for Pattern Recognition. The modification indices for this test indicated that constraining the errors for sessions 5 and 6 to be uncorrelated was the single greatest source of misfit between the data and the model. Therefore, a model which assumed that these errors were correlated and that the correlation was equal in the two samples was fitted to the data. This model substantially improved the fit of the model to the data (chi-square = 99.89; change in chi-square = 28.00, 1 df, $p < .001$). The ATLI also increased relative to the model which assumed the errors for these two sessions were uncorrelated (.794 versus .761). The modification index for the correlated error term was small in both samples, so it was reasonable to accept the hypothesis that the covariation between the errors was equal in the two samples.

Measurement Precision for Different Sessions

The general applicability of a model with constant true score variance coupled with constant error variance after the first test session makes it reasonable to examine the parameter estimates for this model. A key question in this examination is how the measurement precision of the tests changes from the first to later sessions. The typical correlation between scores from the initial session and scores from later sessions tends to be lower than the average correlation among the scores from later sessions. This observation could be explained by a larger error in

measurement for the first session, and the results indicated that the measurement error actually was larger for the first session for 16 of 18 models tested (Table 6).

Table 6 illustrates several interesting points. First, the absolute reliability of the tests generally was modest. Applying Lord and Novick's (1968) definition of reliability as the proportion of true score variance relative to total variance, reliability estimates ranged from .169 to .729 for the initial session and from .469 to .795 for later sessions. If Nunnally's (1978) rule of thumb that reliability should be at least .60 were applied, 9 of 18 measures would not be acceptable for the initial session, and 6 of 18 would not be acceptable after the second trial. The number of errors and percent correct were particularly unreliable with values remaining below .60 even after the first trial for the Manikin test and the Sternberg Memory test.

Table 6
Reliability Comparisons between Session 1 and Later Sessions

		Epsilon for Session:		Reliability for Session:	
	<u>Lambda</u>	<u>1</u>	<u>Later</u>	<u>1</u>	<u>Later</u>
<u>Pattern Recognition</u>					
Net Correct	4.39	15.25	10.10	.558	.656
# Correct	7.16	75.29	31.40	.405	.620
# Errors	7.30	19.79	24.22	.729	.687
% Correct	4.01	6.67	7.06	.707	.695
Avg. Latency	.037	.001	.001	.578	.578
SD Latency	.022	.001	.000	.326	1.000
<u>Manikin</u>					
Net Correct	3.90	7.82	7.76	.660	.662
# Correct	12.46	74.55	53.53	.675	.744
# Errors	1.74	7.74	3.43	.281	.469
% Correct	2.078	21.24	4.48	.169	.491
Avg. Latency	.249	.027	.019	.697	.795
SD Latency	.111	.014	.008	.468	.606
<u>Sternberg</u>					
Net Correct	4.42	7.77	7.71	.715	.717
# Correct	14.36	78.85	65.03	.723	.760
# Errors	2.17	8.84	4.64	.348	.505
% Correct	2.85	21.09	9.89	.278	.451
Avg. Latency	.276	.051	.045	.599	.629
SD Latency	.326	.086	.070	.553	.603

NOTE: "Net Correct" models were computed using the number of correct responses minus the number of incorrect responses divided by the length of test in minutes. Reliability was computed as $(\text{Lambda}^2)/(\text{Lambda}^2 + \text{Epsilon})$.

Discussion

This study produced three findings that extend the empirical bases for designing efficient field studies of cognitive functioning. First, the prior finding that cognitive assessment battery tests typically have constant true score variance across an entire series of sessions, even the first session, was replicated. Provided this true score variance reflects differences in cognitive functioning pertinent to the proposed interpretation of the test scores, this finding means that the tests are valid from the first administration onward. Therefore, valid inferences can be drawn from studies involving even just a single measurement of the cognitive functions. This observation can increase the efficiency of field research by permitting the use of research designs involving only one or two testing sessions in place of designs involving more extensive pre-testing. This conclusion applies provided the research design controls other factors which could invalidate the inferences, such as failure to randomly select and assign subjects to experimental and control groups or a lack of counterbalancing to control for learning curve effects.

The second important finding was that the estimated error variances of the cognitive tests in the present sample were comparable to those in the Kennedy et al. (1985) sample where direct comparisons were possible. If this result generalizes to other cognitive tests and other samples, each test has a fixed standard error of measurement which can be used in power analysis computations to determine the appropriate sample size for studies involving repeated measures (Cohen, 1969). Typical reliability coefficients are composites of true score variance and error variance (Lord & Novick, 1968). The magnitude of this composite will change if either the error of measurement or the true score variance changes, so reliability is not a direct index of measurement precision. In a repeated measures design, if sources of variance other than the treatment effect and error are held constant, the statistical power of the research design will depend on the magnitude of the treatment effect relative to the error variance. Using total variance will lead to an inflated sample size estimate because the expected error variance will be overestimated in the computations. Demonstrating constant measurement precision for cognitive tests, therefore, is a step toward increasing the efficiency of field research designs by decreasing the sample sizes employed while maintaining satisfactory sensitivity to effects of situational demands where effects actually exist.

The third major finding was that a familiarization session reduced measurement error to a value that remained essentially constant thereafter. Efficient research designs, therefore, will require at least one familiarization session. The effect of a familiarization session can be seen by comparing the error term for significance tests without a familiarization session to the error term for significance tests with a familiarization session. Suppose that a study compared two groups, one of which was exposed to an environmental stressor and one of which was not. If cognitive tests are administered only once during or after exposure, the estimated error variance for the groups would consist of the variance in true individual differences with respect to the cognitive abilities measured plus the variance arising from errors in measurement. In terms of the results presented in Table 6, this nominal error variance would be equal to the sum of the square of the lambda value for a performance measure plus the first session error variance. If a familiarization session is conducted prior to collecting experimental data, the scores from this test should be used as a covariate in the analysis of performance during or after exposure. This use of the familiarization scores is legitimate because available evidence suggests these scores are valid assessments of individual differences in cognitive abilities. Given this covariate, the error variance for between-group comparisons will be equal to the estimated error variance (epsilon) for the later sessions in Table 6. The increase in efficiency resulting from a familiarization session can be estimated by comparing this error variance to the nominal error variance in the single session research design. Based on Table 6, a familiarization session will lead to error variance estimates roughly 2.5 to 5 times less than would be obtained without familiarization.

The increase in research design efficiency resulting from the use of a familiarization session can be illustrated by contrasting the effect of such a session with competing strategies that could be considered as means of increasing the power of a research design. The primary alternative in this case would be an increase in sample size. If no familiarization session were conducted, the sampling variance for the error term in the comparisons between groups would decrease linearly with sample size. This rate of change implies that sample size would have to be 2.5 to 5 times as large to achieve the same gains in precision as those provided by a familiarization session. The time required for testing would be 2.5 to 5 times greater because this time would be directly determined by the number of subjects to be tested. A familiarization session would

double the time required to test each subject, so the total testing time with a familiarization session would be equal to that if twice as many subjects were tested once each. The total testing time required for a research design involving a single testing session per subject therefore would be 25% ($2.5/2 = 1.25$) to 150% ($5/2 = 2.50$) greater than that required in a design with a familiarization session.

Several qualifications of the above generalizations should be considered in the design of field studies. First, although simple research designs requiring only one or two testing sessions can provide valid empirical bases for inference, more extensive testing still retain the advantages associated with aggregation of multiple observations whenever more data collection sessions are feasible. Second, the fact that cognitive tests have stable psychometric characteristics from the second session onward is only one point to consider in developing valid research designs. Additional design attention must be directed to problems such as ensuring either random assignment to conditions in the case of a single trial design and controlling for learning curve effects (AGARD Working Group No. 12, 1989). Third, it is important to note the conditional definition of valid variance employed in this study. It was assumed that the stable true score component of the tests represented valid variance, but this assumption must be justified for a given performance measure by embedding it in a theory which provides a context for interpreting the scores. This is a significant concern for cognitive test battery users because the development of test batteries has proceeded in parallel with substantial developments in cognitive psychology. These developments have not yet been fully integrated into the design of test batteries and the configurations of test protocols undertaken with these batteries. Finally, the number of tests and performance measures within tests which have been considered is limited. This statement applies even when the present findings are combined with the prior results obtained by Vickers and Kusulas (1992). Accordingly, caution is appropriate when generalizing from these findings to the full range of cognitive tests available to field researchers. Within the constraints imposed by these caveats, the present results provide a basis for improving the efficiency of field research designs.

References

- AGARD Working Group 12. (1989). Human performance assessment methods. Loughton, Essex, UK: NATO Advisory Group for Aerospace Research and Development, AGARDograph No. 308.
- Baddeley, A. (1986). Working memory. Oxford: Clarendon Press.
- Benson, A. J., & Gedy, J. L. (1983). Logical processes in the resolution of orientation conflict. Report 259. Farnborough, UK: Royal Air Force Institute of Medicine.
- Bentler, P. M., & Mooijart, A. (1989). Choice of structural model via parsimony: A rationale based on precision. Psychological Bulletin, 106, 315-317.
- Bittner, A. C., Jr., Carter, R. C., Kennedy, R. S., Harbeson, M. M., & Krause, M. (1986). Performance Evaluation Tests for Environmental Research (PETER): Evaluation of 114 measures. Perceptual and Motor Skills, 63, 683-709.
- Carter, R., & Woldstad, J. (1985). Repeated measurements of spatial ability with the manikin test. Human Factors, 27, 209-219.
- Cohen, J. (1969). Statistical power analyses for the social sciences. NY: Academic Press.
- Englund, C.E., Reeves, D. L., Shingledecker, C. A., Thorne, D. R., Wilson, K. P., & Hegge, F. W. (1985). Unified Tri-Service Cognitive Performance Assessment Battery (UTC-PAB). I Design and specification of the battery. Report No. 87-10. San Diego, CA: Naval Health Research Center.
- Glymour, C., Scheines, C., Spirtes, P., & Kelly, K. (1987). Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling. NY: Academic Press.
- Jones, M. B., Kennedy, R. S., & Bittner, A. C., Jr. (1981). A video game for performance testing. American Journal of Psychology, 94, 143-152.
- Joreskog, K. G., & Sorbom, D. (1981). LISREL: Analysis of linear structural relationships by the method of maximum likelihood. Chicago: International Educational Services.
- Kennedy, R. S., Dunlap, W. P., Jones, M. B., Lane, N. E., & Wilkes, R. L. (1985). Portable human assessment battery: Stability, reliability, factor structure, and correlation with tests of intelligence. Orlando, FL: Essex Corporation.
- Kivii, H., & Speed, T. P. (1982). Structural analysis of multivariate data: A review. In S. Leinhardt (ed.), Sociological Methodology, 1982. (pp. 209-289). San Francisco: Jossey-Bass.

- Klein, R., & Armitage, R. (1979). Rhythms in human performance: 1 1/2-hour oscillations in cognitive style. Science, 204, 1326-1328.
- Lord, F.M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.
- Meehl, P. E. (1991). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant using it. Psychological Inquiry, 1, 108-141.
- Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stillwell, C. D. (1989). Evaluation of goodness of fit indices for structural equation models. Psychological Bulletin, 105, 430-455.
- Nunnally, J. C. (1978). Psychometric theory. NY: McGraw-Hill.
- Perez, W. A., Masline, P. J., Ramsey, E.G., & Urban, K. E. (1987). Unified tri-services cognitive performance assessment battery: Review and methodology. (AAMRL-TR-87-007). Wright-Patterson Air Force Base, OH: Armstrong Aerospace Medical Research Laboratory.
- Sternberg, S. (1966). High-speed scanning in human memory. Science, 153, 652-654.
- Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.
- Vickers, R. R., Jr., & Kusulas, J. W. (1992). Field applications of cognitive assessment batteries: Initial tests of alternative measurement models (Technical Report No. 92-8). San Diego, CA: Naval Health Research Center.

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE April 1991		3. REPORT TYPE AND DATE COVERED Interim 1 Jan-30 Sep 1991
4. TITLE AND SUBTITLE Replication and Extension of Measurement Models for the Manikin Test, Sternberg Memory Test, and Pattern Recognition Test		5. FUNDING NUMBERS Program Element: 62233N Work Unit Number: M33B30.001-6104		
6. AUTHOR(S) Vickers, Jr., R.R., Hilbert, R., Hodgdon, J.A., Hesslink, R.L., & Hackney, A.C.		8. PERFORMING ORGANIZATION Report No. 92-13		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Health Research Center P. O. Box 85122 San Diego, CA 92186-5122		10. SPONSORING/MONITORING AGENCY REPORT NUMBER		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Naval Medical Research and Development Command National Naval Medical Center Building 1, Tower 2 Bethesda, MD 20889-5044		11. SUPPLEMENTARY NOTES		
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.		12b. DISTRIBUTION CODE		
13. ABSTRACT (Maximum 200 words). Computerized cognitive assessment batteries provide potentially valuable methods of assessing cognitive performance in field settings, but it often is difficult to follow recommendations for extensive baseline testing in these settings. The performance of male Norwegian military personnel (n = 36) on computerized tests of pattern recognition, memory, and spatial orientation tasks was studied to replicate and extend prior findings suggesting that valid measurements of cognitive functions can be obtained even when these recommendations cannot be followed. Structural equation models replicated prior findings that these cognitive tests produce performance measures with constant true score variance from the first session onward and constant error variance from the second session onward. Cross-validation of previously developed structural models indicated comparable error variances across samples, but differences in true score variance were observed which presumably represented sampling variance in the underlying cognitive abilities. These cognitive tests can produce valid results in field studies even when baseline sessions must be limited or omitted if the experimental design and analysis procedures control for other threats to validity (e.g. learning effects).				
14. SUBJECT TERMS cognitive performance memory pattern recognition		spatial orientation military personnel structural equations		15. NUMBER OF PAGES 25
17. SECURITY CLASSIFICATION OF REPORT Unclassified		18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified		16. PRICE CODE
19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified		20. LIMITATION OF ABSTRACT Unlimited		